Extrinsic evaluation of sentence alignment systems

Abdul-Rauf, Sadaf ; Fishel, Mark ; Lambert, Patrik ; Noubours, Sandra ; Sennrich, Rico

Abstract: Parallel corpora are usually a collection of documents which are translations of each other. To be useful in NLP applications such as word alignment or machine translation, they first have to be aligned at the sentence level. This paper is a user study briefly reviewing several sentence aligners and evaluating them based on the performance achieved by the SMT systems trained on their output. We conducted experiments on two language pairs and showed that using a more advanced sentence alignment algorithm may yield gains of 0.5 to 1 BLEU points.

Originally published at:
Abdul-Rauf, Sadaf; Fishel, Mark; Lambert, Patrik; Noubours, Sandra; Sennrich, Rico (2012). Extrinsic evaluation of sentence alignment systems. In: Workshop on Creating Cross-language Resources for Disconnected Languages and Styles, Istanbul, 27 May 2012, 6-10.

Extrinsic Evaluation of Sentence Alignment Systems

Sadaf Abdul-Rauf[1], Mark Fishel[2], Patrik Lambert[1], Sandra Noubours[3], Rico Sennrich[2]

[1]LIUM, University of Le Mans, France
sadaf.abdul-rauf,patrik.lambert@lium.univ-lemans.fr

[2]Institute of Computational Linguistics, University of Zurich, Switzerland
sennrich,fishel@cl.uzh.ch

[3]Fraunhofer FKIE, Wachtberg, Germany
sandra.noubours@fkie.fraunhofer.de

Abstract

Parallel corpora are usually a collection of *documents* which are translations of each other. To be useful in NLP applications such as word alignment or machine translation, they first have to be aligned at the sentence level. This paper is a user study briefly reviewing several sentence aligners and evaluating them based on the performance achieved by the SMT systems trained on their output. We conducted experiments on two language pairs and showed that using a more advanced sentence alignment algorithm may yield gains of 0.5 to 1 BLEU points.

Keywords: sentence alignment, parallel corpora, evaluation

1. Introduction

Parallel corpora[1] constitute an essential cross-language resource whose scarcity for a given language pair and domain restricts the development of data-driven natural language processing (NLP) approaches for that language pair and domain. In this respect, building a parallel corpus helps connecting the considered languages.

After collection, the size of the translated segments forming the parallel corpus are usually of the order of entire documents (e.g. European Parliament sessions or newspaper articles). Learning word correspondences with this kind of examples is an ambiguous task. The ambiguity may be reduced by first decreasing the size of the segments within each pair. This task is called sentence alignment and consists of finding correspondences between segments such as sentences or small paragraphs within a pair of translated documents. The existence of (meta-)textual information such as time stamps (subtitles), speaker information (Europarl (Koehn, 2005)), or paragraphs/chapters/smaller documents provides anchors at which the two sides are certainly aligned. It may thus considerably reduce the complexity of the sentence alignment task. The more fine-grained we can align the text based on textual structure, the easier sentence alignment becomes.

This paper details a user study initiated at the 5th Machine Translation Marathon[2], and whose aim was to evaluate sentence alignment tools on different types of document-aligned parallel corpora and measure its impact on an NLP task, namely Statistical Machine Translation (SMT). The test was conducted on two language pairs. First, on NIST 2008 Urdu–English training data, which contains documents of about 17 sentences in average, with no informative meta- or textual information. Second, on the concatenation of three collections of French–English texts:

- the BAF corpus,[3] composed of very long documents (thousands of lines) with few possible anchors in the text.
- the News Commentary corpus, a corpus of news commentary articles crawled from the web[4], with HTML paragraph mark-up information.
- a corpus crawled from Rapid[5], a site with press releases of the European Union (also containing paragraph mark-up information)

We evaluated five unsupervised sentence alignment tools: the Gale and Church algorithm, Microsoft's Bilingual Sentence Aligner (MBA), Hunalign, Gargantua and Bleualign. In the next section, we describe these sentence alignment tools. Then we present experimental results obtained on the Urdu–English and French–English data. Finally, we draw some conclusions.

2. Sentence Alignment Tools

All five sentence alignment tools that we evaluated use a dynamic programming search to find the best path of sentence pairs through a parallel text. This means that all of them assume that the texts are ordered monotonically and none of the tools is able to extract crossing sentence pairs. For texts with major changes in sentence order between two language version, parallel sentence extraction may be preferrable to searching a global sentence alignment (Fung and Cheung, 2004). All tools also resort to some pruning strategy to restrict the search space.

[1]A parallel corpus is a collection of segment pairs, the two segments within each pair being translation of each other.

[2]http://lium3.univ-lemans.fr/mtmarathon2010/

[3]http://rali.iro.umontreal.ca/Ressources/BAF/

[4]http://www.project-syndicate.org/

[5]http://europa.eu/rapid

While some of the tools support the use of external resources (i.e. bilingual dictionaries in the case of Hunalign, and existing MT systems for Bleualign), all systems learned their respective models from the parallel text itself.

2.1. Gale and Church Algorithm

The Gale and Church (1991; 1993) algorithm is based on character based sentence length correlations, i.e. the algorithm tries to match sentences of similar length and merges sentences, if necessary, based on the number of words in the sentences. The alignment model proposed by Gale and Church (1993) makes use of the fact that longer/shorter sentences in one language tend to be translated into longer/shorter sentences in the other. A probabilistic score is assigned to each proposed sentence pair, based on the sentence length ratio of the two sentences (in characters) and the variance of this ratio. This probabilistic score is then used in the dynamic programming framework to get the maximum likelihood alignment of sentences. Some corpora aligned using this algorithm include the Europarl corpus (Koehn, 2005) and the JRC-Acquis (Steinberger et al., 2006) among others.

2.2. Bilingual Sentence Aligner (MBA)

The Bilingual Sentence Aligner (Moore, 2002) combines a sentence-length-based method with a word-correspondence-based method. While sentence alignment based on sentence-length is relatively fast, lexical methods are generally more accurate but slower. Moore's hybrid approach aims at realising an accurate and computationally efficient sentence alignment model that is not dependent on any additional linguistic resources or knowledge.

The aligner implements a two-stage approach. First the corpus is aligned based on sentence length. The sentence pairs that are assigned the highest probability of alignment are then used as training data for the next stage. In this second stage, a lexical model is trained, which is a modified version of IBM model 1. The final alignment model for the corpus combines the initial alignment model with IBM model 1. These alignments are therefore based on both sentence length and word correspondences and comprise 1-to-1 correspondences with high precision.

2.3. Hunalign

Hunalign (Varga et al., 2005) implements an alignment algorithm based on both sentence length and lexical similarity. It is thus in general similar to Moore's algorithm. The main difference is that Hunalign uses a crude word-by-word dictionary-based replacement instead of IBM model 1. On one hand this results in significant speed gains. More importantly, however, it provides flexible dependence on the dictionary, which can be pre-specified (if one is available) or learned empirically from the data itself.

In case a dictionary is not available, an initial pass is made, based only on sentence length similarity, after which the dictionary is estimated from this initial alignment and a second pass, this time with the dictionary is made.

Although Hunalign is optimised for speed, its memory consumption is its weak spot; in reality it cannot handle parallel corpora larger than 20 thousand sentences – these have to be split into smaller chunks, which results in worse dictionary estimates.

Language	Docs	Max Len.	Ave. Len.	Segm.	Words
Urdu	5282	1003	17.7	93 332	1800 k
English	5282	878	16.9	89 323	2027 k
French	3461	7077	54.2	187 656	4104 k
English	3461	6890	54.1	187 213	3486 k

Table 1: Statistics for the training data set for NIST Urdu–English data and for the French–English data (k stands for thousands).

2.4. Gargantua

Gargantua (Braune and Fraser, 2010) aims to improve on the alignment algorithm by Moore (2002) by replacing the second pass of Moore's algorithm with a two-step clustering approach. As in Moore's algorithm, the first pass is based on sentence-length statistics and used to train an IBM model. The second pass, which uses the lexical model from the first pass, consists of two steps. In a first step, a sequence of 1-to-1 alignments is obtained through dynamic programming. In a second step, these are merged with unaligned sentences to build 1-to-many and many-to-1 alignments.

2.5. Bleualign

Bleualign (Sennrich and Volk, 2010) uses an automatic translation of the source text as an intermediary between the source text and the target text. A first alignment is computed between the translated source text and the target text by measuring surface similarity between all sentence pairs, using a variant of BLEU, then finding a path of 1-to-1 alignments that maximises the total score through dynamic programming. In a second pass, further 1-to-1, many-to-1 and 1-to-many alignments are added through various heuristics, using the alignments of the first pass as anchors.

Bleualign does not build its own translation model for the translation of the source text, but requires an external MT system. In order not to skew the evaluation by using additional resources, we followed Sennrich and Volk (2011) in performing a bootstrapped alignment. As a first step, we aligned the parallel text with the Gale & Church algorithm. Then, we built a SMT system out of this aligned parallel text, and automatically translated the (unaligned) source text. This translation is the basis for the final alignment with Bleualign.

3. Experiments

The aim of the study was to use each sentence aligner to find correspondences at the sentence level in a document-aligned parallel corpus. Then an SMT system was trained from the resulting sentence-aligned parallel corpus, tuned on a development set and used to translate a test set. The sentence aligners were evaluated based on the quality of the translation with respect to automated metrics. The experiment was conducted on two language pairs. The statistics of the document-aligned training data for each language

Set	Language	Segments	Words	Vocabulary	Lmean	Ref.
Dev.	Urdu	923	28.1 k	5.4 k	30.3	1
1st ref.	English	923	24.2 k	5.0 k	26.3	
Test	Urdu	1862	42.3 k	6.5 k	22.7	4
1st ref.	English	1862	38.2 k	6.2 k	20.5	
Dev.	French	2051	55.4 k	9.2 k	27.0	1
1st ref.	English	2051	49.8 k	8.4 k	24.3	
Test	French	2525	72.5 k	11.2 k	28.7	1
1st ref.	English	2525	65.6 k	9.7 k	26.0	

Table 2: Basic statistics for the translation system development and test data sets (k stands for thousands, Lmean refers to the average segment length in number of words, and Ref. to the number of available translation references).

pair are presented in Table 1. These statistics are the number of documents, the maximum document length and the average document length in segments, the total number of segments and the total number of running words in the corpus. The statistics of the development and test data for the SMT systems are presented in Table 2. The statistics shown are the number of segments, the number of words, the vocabulary size (or number of distinct words), the average segment length in number of words and the number of available translation references.

3.1. Urdu–English Task

The Urdu–English data presented in Tables 1 and 2 were provided at NIST 2008 Machine Translation evaluation.[6] The available parallel training and development corpora were only aligned at the document level. We used the training data for the unsupervised sentence alignment. We aligned a part of the development data at the sentence level with the Bleualign tool to build a corpus to tune the SMT systems (Urdu "Dev." in Table 2). Our test set for extrinsic evaluation was the official NIST 2008 test set (Urdu "Test" in Table 2).

The output of the sentence aligners contains at most the same number of tokens as in the training corpus. For some segments, they indeed fail to find any corresponding segment in the other side of the corpus. Table 3 indicates the coverage in terms of number of tokens achieved by the various aligners tested. The % columns indicate the percentage of tokens in the sentence aligned parallel texts compared to the original amount in the training corpus. Gale and Church, Gargantua and Hunalign achieved a coverage around 95%. Bleualign achieved a slightly lower coverage (close to 90%). The MBA only output less than 45% of the input tokens. This can be explained by two reasons. First, it was used with its default precision threshold, which was particularly selective because the Urdu–English data may be noisy or not strictly parallel. A different threshold could have allowed the tool to achieve a higher coverage. Second, the MBA can only extract 1-to-1 correspondences.

The parallel texts described in Table 3 were used to train phrase-based SMT systems with the Moses toolkit (Koehn et al., 2007). In order to stick to the tight MT Marathon schedule, we used an existing language model, trained with news data and data from the European Parliament and the United Nation proceedings.[7] Thus the target side of the sentence-aligned training corpus may not be included in the language model training data. Table 4 shows the scores of three automated MT metrics, namely BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006), obtained by the SMT system trained on the output of each sentence aligner. The evaluation was case-sensitive. The values shown are the average and standard deviation over 3 MERT runs with different random seeds. The values in bold are possibly the best one taking the error range into account.

	Segments (k)		Tokens (k)			
	Urdu	English	Urdu	%	English	%
Training	93.3	89.3	2027	100.0	1800	100.0
Bleualign	65.6		1821	89.9	1607	89.3
Gale&Church	70.0		1925	95.0	1729	96.1
Gargantua	71.1		1943	95.9	1737	96.5
Hunalign	68.7		1950	96.2	1670	92.8
MBA	40.3		902	44.5	745	41.4

Table 3: Coverage on Urdu–English data

Aligner	BLEU	METEOR	TER
Bleualign	**18.1 ±0.3**	**36.0 ±0.2**	**67.9 ±0.7**
Gale&Church	17.0 ±0.3	**35.6 ±0.7**	70.8 ±1.0
Gargantua	**18.1 ±0.2**	**35.6 ±0.4**	**68.1 ±0.7**
Hunalign	17.1 ±0.4	35.3 ±0.2	**69.5 ±1.4**
MBA	17.2 ±0.2	35.4 ±0.2	70.9 ±0.8

Table 4: SMT results on Urdu–English data.

Bleualign and Gargantua tools achieved the highest rank according to all three metrics. Gale and Church and Hunalign methods ranked first according to only one metric. With the corresponding SMT system trained on half the data, MBA achieved worse scores than the other tools according to all metrics. However, the relative difference was below 5%. Still, on this data set one can achieve a significant performance gain by using one of the best tools versus using one of the most basic ones (about 1 BLEU point, 0.5 Meteor point and more than 1.5 TER point).

[6] http://www.itl.nist.gov/iad/mig/tests/mt/2008/

[7] These data are available at http://www.statmt.org/wmt10/.

3.2. French–English Task

We repeated our study on the French–English data, whose statistics are presented in Tables 1 and 2. The training corpus for sentence alignment was described in Sect. 1. The development (French "Dev." in Table 2) and test data (French "Test" in Table 2) were respectively the test set of the 2008 and 2009 Workshop of Statistical Machine Translation shared tasks (see footnote 7).

Table 5 indicates the coverage achieved by the various aligners tested on the French–English data. With this data set the coverage is higher, and the difference between aligners is lower. In particular, the MBA coverage is only 13% lower than that of the aligner with best coverage.

	Segments (k)		Tokens (k)			
	French	English	French	%	English	%
Training	187.7	187.2	4105	100.0	3487	100.0
Bleualign	140.7		3962	96.5	3392	97.3
Gale&Church	141.6		4022	98.0	3440	98.7
Gargantua	142.4		4005	97.6	3430	98.4
Hunalign	142.4		3996	97.4	3414	97.9
MBA	131.7		3503	85.3	3014	86.4

Table 5: Coverage on French–English data

Table 6 shows the (case-sensitive) scores of automated MT metrics achieved by the SMT systems trained (in the same way as in Sect. 3.1.) on the French–English parallel texts output by the different sentence aligners. On this task

Aligner	BLEU	METEOR	TER
Bleualign	**21.07 ±0.07**	**38.83 ±0.15**	**61.2 ±0.2**
Gale&Church	20.64 ±0.07	38.54 ±0.15	61.7 ±0.2
Gargantua	20.83 ±0.07	38.63 ±0.04	**61.1 ±0.1**
Hunalign	**21.03 ±0.10**	**38.68 ±0.10**	**60.9 ±0.2**
MBA	20.91 ±0.03	**38.85 ±0.14**	61.4 ±0.2

Table 6: SMT results on French–English data.

the difference between aligners is lower than on the Urdu–English task. This may be explained by the presence in a part of the corpus of HTML mark-up information, such as paragraphs, sub-sections or links, which makes the sentence alignment task easier. By using the best aligner instead of the worst one, one can achieve a gain of 0.4 BLEU point, 0.3 Meteor point and 0.5 TER point. Bleualign and Hunalign ranked first according to all three metrics. Gargantua and MBA ranked first according to one metric, and the Gale and Church method did not rank first at all.

4. Concluding Remarks

We carried out a brief review of several sentence aligners and evaluated them on the performance of the SMT systems trained on their output, according to automated MT metrics. The coverage of the sentence aligners, as well as the gain achievable by using the best system, depended on the data set. On our Urdu–English data set, this gain was about 1 BLEU point, 0.5 Meteor point and more than 1.5 TER point. On our French–English data set, this gain was about 0.4 BLEU point, 0.3 Meteor point and 0.5 TER point.

Bleualign was the only tool to be ranked first (taking the error range into account) on both tasks and according to the three metrics computed. Gargantua and Hunalign were ranked first according to all metrics on one task. The Gale and Church and MBA tools were ranked first according to one metric on one task.

Acknowledgments

This work has been partially funded by the French Government under the project COSMAT (ANR ANR-09-CORD-004).

5. References

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. pages 65–72, Ann Arbor, MI, June.

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 81–89, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 57–63, Barcelona, Spain.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, pages 79–86.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 135–144, London, UK. Springer-Verlag.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318, Philadelphia, USA, July.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of AMTA 2010*, Denver, Colorado.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *NODALIDA 2011, Nordic Conference of Computational Linguistics*, May.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, and Dan Tufis. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006*, pages 2142–2147.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP-2005)*, pages 590–596, Borovets, Bulgaria.